**PCT** WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : **G07D 5/00, 7/00, G06K 9/62** A1

(11) International Publication Number: **WO 00/33262**

(43) International Publication Date: 8 June 2000 (08.06.00)

(21) International Application Number: PCT/IB99/02012

(22) International Filing Date: 1 December 1999 (01.12.99)

(30) Priority Data:
9826494.8 2 December 1998 (02.12.98) GB

(71) Applicant *(for all designated States except US)*: MARS, INCORPORATED [US/US]; 6885 Elm Street, McLean, VA 22101-3383 (US).

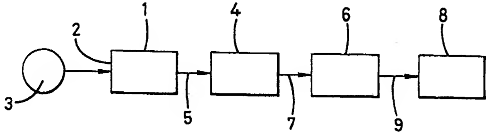
(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: BAUDAT, Gaston [CH/CH]; 74, Grand-Pré, CH-1202 Genève (CH).

(74) Agents: BURKE, Steven, D. et al.; R.G.C. Jenkins & Co., 26 Caxton Street, London SW1H 0RJ (GB).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

**Published**
*With international search report.*
*Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.*

(54) Title: CLASSIFICATION METHOD AND APPARATUS




(57) Abstract

A method of deriving a classification for classifying items of currency into two or more classes comprises measuring known samples for each class, selecting a function corresponding to a non-linear mapping of the feature vector space to a second higher-dimensional space, mapping feature vectors to image vectors, and deriving coefficients representing N-1 axes, where N is the number of classes, in the second space, obtaining values representing the projections of the image vectors for the measured samples onto the N-1 axes, and using those values to derive a separating function for separating the classes equivalent to a function in the second space.

*FOR THE PURPOSES OF INFORMATION ONLY*

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| AL | Albania | ES | Spain | LS | Lesotho | SI | Slovenia |
| AM | Armenia | FI | Finland | LT | Lithuania | SK | Slovakia |
| AT | Austria | FR | France | LU | Luxembourg | SN | Senegal |
| AU | Australia | GA | Gabon | LV | Latvia | SZ | Swaziland |
| AZ | Azerbaijan | GB | United Kingdom | MC | Monaco | TD | Chad |
| BA | Bosnia and Herzegovina | GE | Georgia | MD | Republic of Moldova | TG | Togo |
| BB | Barbados | GH | Ghana | MG | Madagascar | TJ | Tajikistan |
| BE | Belgium | GN | Guinea | MK | The former Yugoslav Republic of Macedonia | TM | Turkmenistan |
| BF | Burkina Faso | GR | Greece | | | TR | Turkey |
| BG | Bulgaria | HU | Hungary | ML | Mali | TT | Trinidad and Tobago |
| BJ | Benin | IE | Ireland | MN | Mongolia | UA | Ukraine |
| BR | Brazil | IL | Israel | MR | Mauritania | UG | Uganda |
| BY | Belarus | IS | Iceland | MW | Malawi | US | United States of America |
| CA | Canada | IT | Italy | MX | Mexico | UZ | Uzbekistan |
| CF | Central African Republic | JP | Japan | NE | Niger | VN | Viet Nam |
| CG | Congo | KE | Kenya | NL | Netherlands | YU | Yugoslavia |
| CH | Switzerland | KG | Kyrgyzstan | NO | Norway | ZW | Zimbabwe |
| CI | Côte d'Ivoire | KP | Democratic People's Republic of Korea | NZ | New Zealand | | |
| CM | Cameroon | | | PL | Poland | | |
| CN | China | KR | Republic of Korea | PT | Portugal | | |
| CU | Cuba | KZ | Kazakstan | RO | Romania | | |
| CZ | Czech Republic | LC | Saint Lucia | RU | Russian Federation | | |
| DE | Germany | LI | Liechtenstein | SD | Sudan | | |
| DK | Denmark | LK | Sri Lanka | SE | Sweden | | |
| EE | Estonia | LR | Liberia | SG | Singapore | | |

# Classification Method and Apparatus

The invention relates to a method and apparatus for classifying items. The invention is concerned especially with the classification of coins or banknotes.

Coins and banknotes inserted into mechanisms, such as vending machines, change machines and the like, are classified, on the one hand according to value, and/or on the other hand, between originals and copies or counterfeits thereof. Various methods of performing such classifications are known. As one example, described in GB 2 238 152 A, the contents of which are incorporated herein by reference. For example, measurements are taken from an inserted coin which represent different features of the coin, such as material and the thickness. Those measurements are then compared with respective stored pairs of values, each set of pair of values corresponding to a respective acceptable denomination of coin. When each measured value falls within the respective range for a given denomination, the inserted coin is classified as belonging to that denomination.

In the type of classification discussed above, the measured values can be regarded as elements in a feature vector, and the acceptable measurements for different denominations correspond to regions in feature space, known as acceptance regions. In the example given above, the feature space is two-dimensional, and acceptance regions are rectangles, but the feature space can

have any number of dimensions, with corresponding complexity in the acceptance regions. For example, GB 2 254 949 A, the contents of which are incorporated herein by reference, describes ellipsoidal acceptance regions in three-dimensional feature space.

Other examples of methods and apparatus for classifying bills and coins are described in EP 0 067 898 A, EP 0 472 192 A, EP 0 165 734 A. Other methods of classification include the use of neural networks, as described, for example, in EP 0 553 402 A and EP 0 671 040 A, the contents of which are also incorporated herein by reference.

A significant problem in the classification of coins is the difficulty of separating different denominations. The population distributions of the different denominations of interest may be such that it is not possible easily to define appropriate acceptance boundaries with which adequately separate the denominations. Another problem is that in order to achieve adequate separation, it may be necessary to consider feature vectors having a large number of elements, which makes it more difficult to understand the various distributions and thus more difficult to obtain suitable acceptance boundaries. These problems are akin to general classification problems in data analysis which has been studied and have led to various different techniques including statistical methods.

As an example of a statistical method of data analysis, principal component analysis ("PCA"), is a method whereby data expressed in one

space is transformed using a linear transformation into a new space, where most of the variation within the data can be explained using fewer dimensions than in the first space. The method of PCA involves finding the eigenvectors and eigenvalues of the covariance matrix of the variables. The eigenvectors are the axes in the new space, with the eigenvector having the highest eigenvalue being the first "principal component" and so on in decreasing size. Details of PCA can be found in textbooks on multivariate analysis, such as "Introduction to Multivariate Analysis" by Chatfield and Collins, see Chapter 4.

Another method of data analysis for classification purposes is linear discriminant analysis ("LDA"). LDA is useful when it is known that the data falls into separate groups. LDA aims to transform the data into a new space so as to maximize the distance between the centre of each group of data as projected onto axes in the new space and also to minimize the variance of each group along the axes. Methods for doing this are described in, for example, "Introduction to Statistical Pattern Recognition" by Fukunaga ("Fukunaga"). In one example, the maximisation is performed by finding a linear transformation which maximises the value of the trace of $C^{-1}V$ where V is the inter-class covariance matrix and C is the covariance matrix of all samples . As explained in Fukunaga, this amounts to finding the eigenvectors and eigenvalues of $C^{-1}V$ . The eigenvectors are the axes of the new space. As

described in the paper, when there are N classes, the new space has N-1 dimensions.

In many situations, neither PCA nor LDA will give adequate separation of the groups of data. A further method of data analysis is non-linear component analysis (NCA), which is based on PCA. In NCA, the data is projected into a new space using a non-linear mapping, and then PCA is performed in the new space. Details of NCA are given in the article "Nonlinear component Analysis as a Kernel Eigenvalue Problem" by Bernhard Scholkopf, Alexander Smola and Klaus-Robert Muller, Neural Computation 10, 1299-1319 (1998). ("Scholkopf".)

A problem with NCA is that the dimension of the non-linear space may be very large, and so the number of principal components is also very large. For a given problem, it is not known how many principal components are needed for a good classification.

Generally, the invention relates to a method of deriving a classification for classifying items of currency comprising measuring known samples for each class and deriving features vectors from the measured samples, mapping the feature vectors to a second space in which there is a clearer separation of the different classes and deriving a separating function using the separation in the second space.

More specifically, the present invention provides a method of deriving a classifier for classifying items of currency into two or more classes

comprising measuring known samples for each class and deriving feature vectors from the measured samples, selecting a function corresponding to a mapping of the feature vector space to a second space, mapping feature vectors to image vectors, and deriving coefficients representing N-1 axes, where N is the number of classes, in the second space, obtaining values representing the projections of the image vectors for the measured samples onto the N-1 axes, and using those values to derive a separating function for separating the classes equivalent to a separating function in the second space.

The invention also provides a method for classifying an item of currency comprising measuring features of the item, generating a feature vector from the measured values, and classifying the item using a classifying derived by a method according to any one of claims 1 to 6.

The invention also provides an apparatus for classifying items of currency comprising measuring means for measuring features of an item of currency, feature vector generating means for generating a feature vector from the measured values, and classifying means for classifying the item using a classifier derived according to the method of any one of claims 1 to 6.

The invention also provides an apparatus for classifying items of currency comprising measuring means for measuring features of an item of currency, feature vector generating means for generating a feature vector from the measured values, and classifying means for classifying the item using a function corresponding to a non-linear mapping of the feature vector space to

a second higher-dimensional space, mapping feature vectors to image vectors, and coefficients representative of N-1 axes, where N is the number of classes that can be classified by the apparatus, in the second space, and a function equivalent to a separating function in the second space.

An embodiment of the invention will be described with reference to the accompanying drawings of which:

Fig. 1 is a block diagram of a classification system.

Fig. 2 is a graph showing a distribution of coin data; and

Fig. 3 is a graph showing a projection of the data of Fig. 2 onto new axes.

The invention will be described with reference to a coin validator.

In Fig. 1, box 1 designates a measuring system which includes an inlet 2, a transport system in a form of a coin inlet and coin transport path (not shown) for presenting a sample 3 and a sensor system (not shown) for measuring physical quantities of the sample. The measuring system 1 is connected to a processing system 4 by means of a data bus 5. Processing system 4 is connected to a classifier 6 by means of a data bus 7. The output of the classifier 6 is connected to a utilization system 8 by means of a data output bus 9. The utilization system 8 is in this example a vending machine, but may also be, for example, a money exchange machine.

The measuring system 1 measures features of an inserted coin 3. The measured features are assembled into a feature vector having n elements,

where each element corresponds to a measured feature by the processing system 4. In the present example, the sensor system measures values representative of the material, thickness and diameter of an inserted coin, using known techniques (see, for example, GB 2 254 949 A) and those values are the three elements of the corresponding feature vector. Briefly, each sensor comprises one or more coils in a self-oscillating circuit. In the case of the diameter and thickness sensors, a change in the inductance of each coil caused by the proximity of an inserted coin causes the frequency of the oscillator to alter, whereby a digital representation of the respective property of the coin can be derived. In the case of the conductivity sensor, a change in the Q of the coil caused by the proximity of an inserted coin causes the voltage across the coil to alter, whereby a digital output representative of conductivity of the coin may be derived. Although the structure, positioning and orientation of each coil, and the frequency of the voltage applied thereto, are so arranged that the coil provides an output predominantly dependent upon a particular one of the properties of conductivity, diameter and thickness, it will be appreciated that each measurement will be affected to some extent by other coin properties.

Of course, many different features representative of items of currency can be measured and used as the elements of the feature vectors. For example, in the case of a banknote, the measured features can include, for example, the width of the note, the length of the note, and the intensity of

reflected or transmitted light for the whole or part of the note. As an example, a measuring system can be arranged to scan a banknote along N lines using optical sensors. Each scan line contains L individual areas, which are scanned in succession. In each area, there are measurements of M different features. More specifically, for each area, measurements are made of the reflectance intensities of red, green and infra-red radiation. The total number of measurements for a banknote is therefore L x M x N. These measurements form the components of a feature vector for the respective specimen, so that the feature vector has L x M x N components. Alternatively, the measurements can be processed in a different way to obtain a feature vector representative of the measured specimen. For example, local feature vectors for each measured area can be formed made up of the M measurements for that area, so that each local feature vector has M components. The local feature vectors can then be summed over the area of the banknote to obtain an M dimensional feature vector representative of the entire specimen.

The feature vector is then input to the classifier 6. The classifier 6 determines whether the sample belongs to any one of predetermined classes, using the feature vector and predetermined classification criteria including a separating function. If the sample is identified as belonging to an acceptable denomination of banknote, then it is accepted and the corresponding value of the note is credited. If the sample is identified as belonging to a known counterfeit group, it is rejected.

In this example, the system is for classifying two denominations of coins and one known counterfeit. A two-dimensional representation of the distribution in measurement space is shown in Fig. 2. The crosses represent samples of the first denomination, the dots represent counterfeits of the first denomination and the circles represent samples of the second denomination.

The derivation of the separating function will be described below in general terms. The method of classification will then be described, also in general terms, followed by an explanation of the application of the general method to the specific example.

Briefly, a method for deriving a separating function according to an embodiment of the invention maps the input space, that is the space of the measured feature vectors, using a non-linear map, into a higher dimensional space with linear properties. Separating hyperplanes are constructed in the mapped space using training data, using the equivalent of an LDA analysis in the mapped space.

The population distribution of the denominations are analysed as discussed below.

Initially, samples of each of the denominations of interest and each of the known counterfeit are measured and corresponding feature vectors are formed. The feature vectors from the samples, when plotted, for example, on a n-dimensional scatter graph, (where n is the number of measured features) fall roughly into groups, known as clusters. These measured samples are then

used to derive a separating function, as described below. In this example, 50 samples for each denomination and 50 samples of the counterfeit, are used.

Before proceeding further, a general explanation of the notation used is provided.

The input space, that is, the space of feature vectors, is defined as X. $X = \bigcup_{l=1}^{N} X_l$, where N is the number of clusters. The cardinality of subspace $X_l$ is denoted by $n_l$, and the number of elements in X is M. Thus $\sum_{l=1}^{N} n_l = M$. $x^t$ is the transpose of vector x.

In the input space, C is the covariance matrix, and

$$C = \frac{1}{M} \sum_{j=1}^{M} x_j x_j^t \qquad (1)$$

The method of the invention uses a kernel function k defining a dot product in a mapped space. Suppose $\phi$ is a non-linear function mapping X into a Hilbert space F.

$$\phi : X \rightarrow F$$

$$x \rightarrow \phi(x) \qquad (2)$$

and $k(x,y) = \phi(x) \bullet \phi(y) = \phi^t(x)\, \phi(y)$

As will be clear from the following discussion, it is not necessary explicitly to construct $\phi$ for a given k, although it can be shown, by Mercer's theorem, if for any k is a continuous kernel of a positive integral operator

which is positive, then a $\phi$ exists (see Schölkopf Section 3 and Appendix C). Nor is it necessary to perform dot products explicitly in F, which may be an infinite dimensional space.

In F, V is the covariance matrix, and

$$V = \frac{1}{M}\sum_{j=1}^{M}\phi(x_j)\phi'(x_j) \qquad (3)$$

We assume that the observations are centred in F, that is, that $\sum_{j=1}^{M}\phi(x_j) = 0$. A method of centering data will be described later.

B is the covariance matrix of the cluster centres, and

$$B = \frac{1}{M}\sum_{l=1}^{N} n_l \overline{\phi_l \phi_l'} \qquad (4)$$

where $\overline{\phi_l}$ is the mean value of the cluster l, that is

$$\overline{\phi_l} = \frac{1}{n_l}\sum_{k=1}^{n_l}\phi(x_{lk}) \qquad (5)$$

where $x_{lj}$ is the element j of the cluster l.

B represents the inter-cluster inertia in F.

V can also be expressed using the clusters as

$$V = \frac{1}{M}\sum_{l=1}^{N}\sum_{k=1}^{n_l}\phi(x_{lk})\phi'(x_{lk}) \qquad (6)$$

V represents total inertia in F.

Let $k_{ij} = k(x_i, x_j)$

and $(k_{ij})_{pq} = (\phi^t(x_{pi})\phi(x_{qj}))$

Let K be an (MxM) matrix defined on the cluster elements by $(K_{pq})_{\substack{p=1\ldots N\\ q=1\ldots N}}$

where $(K_{pq})$ is the covariance matrix between cluster $_p$ and cluster $_q$.

$$K = (K_{pq})_{\substack{p=1\ldots N\\ q=1\ldots N}} \text{ where } K_{pq} = (k_{ij})_{\substack{i\ldots n_p\\ j\ldots n_q}} \qquad (8)$$

$K_{pq}$ is a $(n_p \times n_q)$ matrix

and K is symmetric so that $K_{pq}^t = K_{pq}$

W is the matrix centre, and

$$W = (W_l)_{l=1\ldots N} \qquad (9)$$

where $W_l$ is a $(n_l \times n_l)$ matrix with all terms equal to $\frac{1}{n_l}$.

W is a M x M block diagonal matrix.

The method essentially performs linear discriminant analysis in the mapped space F to maximise inter-cluster inertia and minimise the intra-cluster inertia. This is equivalent to eigenvalue resolution, as shown in Fukunaga. A suitable separating function can then be derived.

More specifically, the method involves finding the eigenvalues λ and eigenvectors v that satisfy

$$\lambda V v = B v \qquad (10)$$

The eigenvectors are linear combinations of elements of F and so there exist coefficients $\alpha_{pq}$ (p=1...N, q=1...$n_p$) such that

$$v = \sum_{p=1}^{N} \sum_{q=1}^{n_p} \alpha_{pq} \phi(x_{pq}) \qquad (11)$$

The eigenvectors of equation (10) are the same as the eigenvectors of

$$\lambda \phi'(x_{ij}) Vv = \phi'(x_{ij}) Bv \qquad (12)$$

(see Schölkopf).

Using the definitions of K and W, and equations (6) and (11), the left-hand side of (12) can be expressed as follows:

$$Vv = \frac{1}{M} \sum_{l=1}^{N} \sum_{k=1}^{n_l} \phi(x_{lk}) \phi'(x_{lk}) \sum_{p=1}^{N} \sum_{q=1}^{n_p} \alpha_{pq} \phi(x_{pq})$$

$$= \frac{1}{M} \sum_{p=1}^{N} \sum_{q=1}^{n_p} \alpha_{pq} \sum_{l=1}^{N} \sum_{k=1}^{n_l} \phi(x_{lk}) \left[ \phi'(x_{lk}) \ \phi(x_{pq}) \right]$$

and $$\lambda \phi'(x_{ij}) Vv = \frac{\lambda}{M} \sum_{p=1}^{N} \sum_{q=1}^{n_p} \alpha_{pq} \phi'(x_{ij}) \sum_{l=1}^{N} \sum_{k=1}^{n_l} \phi(x_{lk}) \left[ \phi'(x_{lk}) \ \phi(x_{pq}) \right]$$

$$= \frac{\lambda}{M} \sum_{p=1}^{N} \sum_{q=1}^{n_p} \alpha_{pq} \sum_{l=1}^{N} \sum_{k=1}^{n_l} \left[ \phi'(x_{ij}) \phi(x_{lk}) \right] \left[ \phi'(x_{lk}) \ \phi(x_{pq}) \right]$$

Using this formulate for all clusters i and for all elements j we obtain:

$$\lambda (\phi'(x_{11}), \ldots, \phi'(x_{1n_q}), \ldots, \phi'(x_{ij}), \ldots, \phi'(x_{N1}), \ldots, \phi'(x_{Nn_N})) Vv = \frac{\lambda}{M} KK\alpha$$

where $\alpha = (\alpha_{pq})_{\substack{p = 1 \ldots N \\ q = 1 \ldots n_p}}$

$= (\alpha_p)_{p = 1 \ldots N}$

where $\alpha_p = (\alpha_{pq}) q = 1 \ldots n_p$

Using equations (4), (5) and (11), for the right term of (14):

$$Bv = \frac{1}{M} \sum_{p=1}^{N} \sum_{q=1}^{n_p} \alpha_{pq} \phi(x_{pq}) \sum_{l=1}^{N} n_l \left[ \frac{1}{n_l} \sum_{k=1}^{n_l} \phi(x_{lk}) \right] \left[ \frac{1}{n_l} \sum_{k=1}^{n_l} \phi(x_{lk}) \right]^t$$

$$= \frac{1}{M} \sum_{p=1}^{N} \sum_{q=1}^{n_p} \alpha_{pq} \sum_{l=1}^{N} \left[ \sum_{k=1}^{n_l} \phi(x_{lk}) \right] \left[ \frac{1}{n_l} \right] \left[ \sum_{k=1}^{n_l} \phi'(x_{lk}) \ \phi(x_{pq}) \right]$$

and $\phi'(x_{ij})Bv=\frac{1}{M}\sum_{p=1}^{N}\sum_{q=1}^{n_p}\alpha_{pq}\sum_{l=1}^{N}\left[\sum_{k=1}^{n_l}\phi'(x_{ij})\phi(x_{lk})\right]\left[\frac{1}{n_l}\right]\left[\sum_{k=1}^{n_l}\phi'(x_{lk})\ \phi(x_{pq})\right]$

For all clusters i and for all elements j we obtain:

$$(\phi'(x_{11}),\ldots,\phi'(x_{1n_1}),\ldots,\phi'(x_{ij}),\ldots,\phi'(x_{N1}),\ldots,\phi'(x_{Nn_N}))Bv=\frac{1}{M}KWK\alpha \qquad (14)$$

Combining (13) and (14) we obtain:

$\lambda KK\alpha = KWK\alpha$

Thus, $\lambda = \dfrac{\alpha' KWK\alpha}{\alpha' KK\alpha}$ (15)

K can be decomposed as K = QR (Wilkinson, 1971) so that $K\alpha = QR\alpha$.

R is upper triangular and Q is orthonormal, that is $Q^tQ = I$.

Q is a Mxr matrix and R is a rxM matrix, where r is the rank of K. It is known that the QR decomposition always exists for a general rectangular matrix.

Then, let $R\alpha = \beta$ (16)

As the rows of R are linearly independent, for a given β, there exists at least one α solution.

Hence $K\alpha = Q\beta$ and $\alpha^tK = \beta^tQ^t$ (K is symmetric).

Substituting in (15)

$\lambda = \dfrac{\alpha' KWK\alpha}{\alpha' KK\alpha}$ (17)

Q is orthonormal so $\lambda\beta = Q^tWQ\beta$ (18)

Equation (18) is in the form of a standard eigenvector equation. As K is singular, the QR decomposition permits work on a subspace Qβ, which simplifies the resolution.

Then the coefficients α can be derived from β from equation (16), and then the eigenvectors from equation (11).

These coefficients α are normalised by requiring that the corresponding vectors v in F be normalised. That is:

$$v^t v = 1 \qquad (19)$$

or (from equation 11)

$$V'V = \sum_{p=1}^{N}\sum_{q=1}^{np}\sum_{l=1}^{N}\sum_{h=1}^{nl}\alpha_{pq}\alpha_{lh}\phi'(x_{pq})\phi(x_{lh}) = 1$$

$$= \sum_{p=1}^{N}\sum_{l=1}^{N}\alpha'_p K_{pl}\alpha_l = 1$$

$$= \alpha' K\alpha$$

$$\text{so}(19) \Rightarrow \alpha' K\alpha = 1 \qquad (20)$$

The steps given above set out how to find the eigenvectors v of equation (10).

As its known from linear discriminant analysis (see, for example, Fukunaga), the number of eigenvectors = N-1 where N is the number of clusters. The image of the clusters in the subspace spanned by the eigenvectors is found by projecting onto the eigenvectors. This is done using the following equation:

for an eigenvector v, and a feature vector x.

$$(\phi^t(x)v) = \sum_{p=1}^{N}\sum_{q=1}^{n_p}\alpha_{pq}\phi^t(x_{pq})\phi(x) = \sum_{p=1}^{N}\sum_{q=1}^{n_p}\alpha_{pq}k(x_{pq},x) \quad (21)$$

As can be seen from the above, the calculation does not require knowledge of $\phi$, or the need to calculate a dot product in F.

It has been shown in experiments that by use of a suitable kernel function, the images of the clusters in the eigenvector subspace are well-separated and, more specifically, may be linearly separable, that is they can be separated by lines, planes or hyperplanes.

Then a suitable separating function can easily be derived for classifying measured articles, using a known technique, such as inspection, averaging, Malalanobis distance, comparison with k nearest neighbours.

As mentioned previously, it was assumed that the observations are centred in F. Centering will now be discussed in more detail. Firstly, for a given observation $x_{ij}$: element j of the cluster i, the image $\phi(x_{ij})$ is centered according to:

$$\tilde{\phi}(x_{ij}) = \phi(x_{ij}) - \frac{1}{M}\sum_{l=1}^{N}\sum_{k=1}^{n_l}\phi(x_{lk}) \quad (22)$$

We have then to define the covariance matrix K with centered points:

$(\tilde{k}_{ij})_{pq} = (\tilde{\phi}(x_{pi}).\tilde{\phi}(x_{qj}))$ for a given cluster p and q.

$$(\tilde{k}_{ij})_{pq} = \left[\phi(x_{pi}) - \frac{1}{M}\sum_{l=1}^{N}\sum_{k=1}^{n_l}\phi(x_{lk})\right]\left[\phi(x_{qj}) - \frac{1}{M}\sum_{h=1}^{N}\sum_{m=1}^{n_m}\phi(x_{hm})\right]$$

$$(\tilde{k}_{ij})_{pq} = (k_{ij})_{pq} - \frac{1}{M}\sum_{l=1}^{N}\sum_{k=1}^{n_l}(1_{ik})_{pl}(k_{kj})_{lq} - \frac{1}{M}\sum_{h=1}^{N}\sum_{m=1}^{n_m}(k_{im})_{ph}(1_{mj})_{hq} + \frac{1}{M^2}\sum_{l=1}^{N}\sum_{k=1}^{n_l}\sum_{h=1}^{N}\sum_{m=1}^{n_m}(1_{ik})_{pl}(k_{km})_{lh}(1_{mj})_{hq}$$

$$\tilde{K}_{pq} = K_{pq} - \frac{1}{M}\sum_{l=1}^{N}1_{pl}K_{lq} - \sum_{h=1}^{N}K_{ph}1_{hq} + \frac{1}{M^2}\sum_{l=1}^{N}\sum_{h=1}^{N}1_{pl}K_{lh}1_{hq}$$

$$\tilde{K} = K - \frac{1}{M}1_N K - \frac{1}{M}K1_N + \frac{1}{M^2}1_N K1_N$$

Where we have introduced the following matrix:

$l_{pl} = (l_{ik})_{i=1,\dots,n_p;k=1,\dots,n_l}$, $(n_p x n_l)$ matrix whose elements are all equal to 1.

$l_N = (l_{pl})_{p=1,\dots,N;l=1,\dots,N}$, $(MxM)$ matrix whose elements are block matrices.

Thus, for non-centred points $\phi(x_{ij})$, we can derive $\tilde{K}$ from K and then solve for the eigenvectors of $\tilde{K}$. Then, for a feature vector x, the projection of the centred $\phi$-image of x onto the eigenvectors $\tilde{v}$ is given by:

$$(\tilde{\phi}^t(x)v) = \sum_{p=1}^{N}\sum_{q=1}^{n_p}\tilde{\alpha}_{pq}\tilde{\phi}^t(x_{pq})\tilde{\phi}(x)$$

The above discussion sets out in general terms the method of general discriminant analysis. The general principles will now be illustrated with reference to the specific example of the coin validator.

Returning to the example of the coin validator at the beginning of the description, the feature vectors each have three elements, and there are three

clusters, corresponding to each of the two denominations of interest and the known counterfeit respectively.

50 samples of each denomination and 50 samples of the counterfeit are input to the measuring system 1. As previously mentioned, the sensor systems measures samples to obtain values representative of the thickness, material and diameter in each case. Corresponding feature vectors are formed from the measured features for each sample.

From the 50 samples feature vectors for each cluster, 37 are randomly selected for use in generating the separating function.

A kernel function is then chosen. The kernel function is chosen on the basis of trial and error so as to choose whichever function gives the best separation results. There are a large number of kernel functions, satisfying Mercer's theorem, which may be suitable. Examples of kernel functions are the polynominal kernel:

$$k\ (x, y) = (x.y)^d;$$

the Gaussian kernel:

$$k\ (x,y) = \exp \frac{(\|x-y\|^2)}{\sigma^2};$$

the hyberbolic tangent kernel:

$$k\ (x,y) = \tanh\ ((x.y)+\theta);\ \text{and}$$

the sigmoid kernel:

$$k(x,y) = (\frac{1}{1+e^{-((x.y)+\theta)}})\ .$$

In this example, the Gaussian kernel is used, with $\sigma^2 = 0.01$.

Using the selected samples and the kernel function, the matrices K and W are calculated. (Equations (8) and (9)).

Then K is decomposed using QR decomposition.

Then eigenvectors β and corresponding eigenvectors are calculated (equation (18)).

Then coefficients α are calculated and normalised (equations (16) and (20)).

Thereafter, the feature vectors of the remaining 13 samples for each cluster are projected onto the eigenvectors v (equation 21) and the results are plotted on a graph for easy inspection. In this example, there are 3 clusters, so there are 2 eigenvectors, and separation is in 2-d space. This is shown in Fig. 3. As can be seen, the clusters are well-separated. More specifically, each cluster is projected on one point, which is the gravity centre. The separation of the projection of the clusters with the eigenvectors is then analysed, and used to derive a separation function. In this example, a linear separating function can easily be derived by inspection. For example, a suitable separating function is:

for eigenvectors $v_1$, $v_2$

and an input vector x

If $\left[(\phi'(x)v_l) > 0 \text{ and } (\phi'(x)v_2) > 0\right]$ then

x belongs to group 1 (that is, it is of the first denomination);

if $\left[(\phi'(x)v_l) > 0 \text{ and} (\phi'(x)v_2) < 0\right]$

then x belongs to group 2 (that is, it is of the second denomination); and

if $\left[(\phi'(x)v_l) < 0 \text{ and} (\phi'(x)v_2) > 0\right]$

then x belongs to group 3 (that is, it is a counterfeit of the first denomination).

Classification for coins of an unknown denomination is then performed as follows. The inserted coin is sensed, and measurements representative of the material, thickness and diameter are obtained, as for the samples. A feature vector is then derived from the measured values. The feature vector is then projected onto the calculated eigenvectors (using equation 21) and the coin is classified in accordance with the projection values and the separating function, as described above.

The analysis of the sample values for the initial data analysis and the derivation of the separating function can be done, for example, using a microprocessor. Similarly, the classifier 6 may be a microprocessor.

As an alternative, the classifier 6 may be a neural network, such as a probabilistic neural network, or a perceptron. For example, the neural network may include N-1 linear output neurones and M hidden neurones, where every kernel computation is a hidden neurone. Then the input weights

are the values $x_{pq}$, and the coefficients $\alpha$ are the weights between the hidden neurones and the output layer.

Also, the classifier may be a linear classifier, or a Support Vector Machine.

The methods of the embodiment described above are equally applicable to a banknote or indeed to a classification of other sorts of items. Other methods of solving (10), for example by decomposing K using eigenvector decomposition, are possible.

In the embodiment, a non-linear mapping to a higher-dimensional space is used. A linear mapping could be used instead. Also, mapping could be to a lower-dimensional space, or to a space of the same dimension as the feature vector space.

**Claims**

1. A method of deriving a classification for classifying items of currency into two or more classes comprising measuring known samples for each class and deriving feature vectors from the measured samples, selecting a function corresponding to a mapping of the feature vector space to a second space, mapping feature vectors to image vectors, and deriving coefficients representing N-1 axes, where N is the number of classes, in the second space, obtaining values representing the projections of the image vectors for the measured samples onto the N-1 axes, and using those values to derive a separating function for separating the classes equivalent to a separating function in the second space.

2. A method as claimed in claim 1 wherein the mapping is a non-linear mapping.

3. A method as claimed in claim 1 or claim 2 wherein the second space is higher-dimensional than the first space.

4. A method as claimed in any one of claims 1 to 3 wherein the coefficients are derived by optimising the separation of the groups of image vectors for each class with respect to the axes.

5. A method as claimed in any one of claims 1 to 4 comprising deriving a matrix V where V is the covariance matrix in the second space and a matrix B where B is the covariance matrix of the class centres in the second space, deriving the solutions to the equation $\lambda Vv=Bv$, and deriving said coefficients from the solutions v.

6. A method as claimed in any one of claims 1 to 5 wherein said function expresses a dot product in the second space in terms of a function on two elements of the feature vector space.

7. A method as claimed in claim 6 wherein said function is k(x,y) where $k(x,y) = (x.y)^d$.

8. A method as claimed in claim 6 wherein said function is k(x,y) where $k(x,y) = \exp\frac{(\|x-y\|^2)}{\sigma^2}$.

9. A method as claimed in claim 6 wherein said function is k(x,y) where $k(x,y) = \tanh((x.y) + \theta)$.

10. A method as claimed in claim 6 wherein said function is k(x,y) where $k(x,y) = (\frac{1}{1 + e^{-((x.y)+\theta)}})$ .

11. A method for classifying an item of currency comprising measuring features of the item, generating a feature vector from the measured values, and classifying the item using a classifying derived by a method according to any one of claims 1 to 10.

12. An apparatus for classifying items of currency comprising measuring means for measuring features of an item of currency, feature vector generating means for generating a feature vector from the measured values, and classifying means for classifying the item using a classifier derived according to the method of any one of claims 1 to 10.

13. An apparatus for classifying items of currency comprising measuring means for measuring features of an item of currency, feature vector generating means for generating a feature vector from the measured values, and classifying means for classifying the item using a function corresponding to a mapping of the feature vector space to a second space, mapping feature vectors to image vectors, and coefficients representative of N-1 axes, where N is the number of classes that can be classified by the apparatus, in the second space, and a function equivalent to a separating function in the second space.

14. An apparatus as claimed in claim 13 wherein the classifying means comprises means for deriving values representing the projection of the image of the feature vector of the measured item onto the or each axis.

15. An apparatus as claimed as claimed in any one of claims 12 to 14 wherein the classifying means comprises a neural network.

16. An apparatus as claimed in any one of claims 12 to 15 comprising a coin inlet and the measuring means comprises sensor means for sensing a coin.

17. An apparatus as claimed in claim 16 wherein the sensor means is for sensing the material and/or the thickness and/or the diameter of a coin.

18. An apparatus as claimed in any one of claims 12 to 14 comprising a banknote inlet and wherein the measuring means comprises sensor means for sensing a banknote.

19. An apparatus as claimed in claim 18 wherein the sensor means is for sensing the intensity of light reflected from and/or transmitted through a banknote.

20. A coin validator comprising an apparatus as claimed in any one of claims 12 to 17.

21. A banknote validator comprising an apparatus as claimed in any one of claims 12 to 14 or 18 or 19.

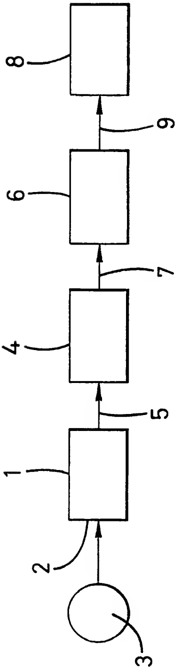
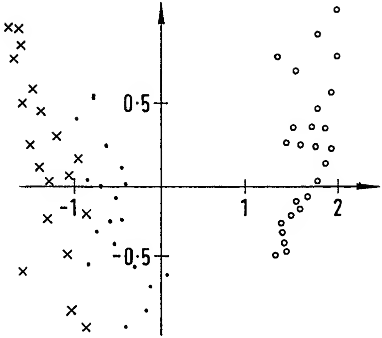



FIG.1

FIG. 2



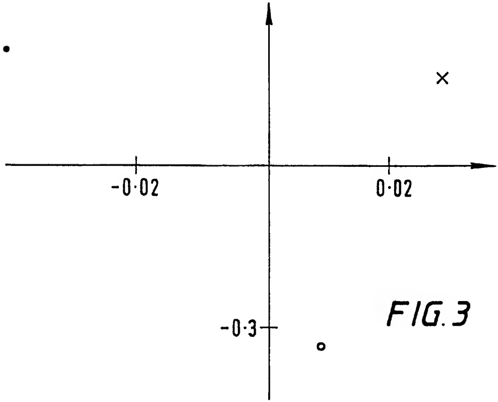




FIG. 3

# INTERNATIONAL SEARCH REPORT

Internat. l Application No
PCT/IB 99/02012

**A. CLASSIFICATION OF SUBJECT MATTER**
IPC 7 G07D5/00 G07D7/00 G06K9/62

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category * | Citation of document, with indication, where appropriate, of the relevant passages | Relevant to claim No. |
|---|---|---|
| X | BURGES C J C: "A tutorial on support vector machines for pattern recognition" DATA MINING AND KNOWLEDGE DISCOVERY, vol. 2, no. 2, 1998, pages 121-167, XP002087854 Chapter 4, 4.1-4.3 | 1-9, 11-13 |
| X | BURGES C J C: "Simplified support vector decision rules" MACHINE LEARNING. PROCEEDINGS OF THE INTERNATIONAL CONFERENCE,XX,XX, 3 July 1996 (1996-07-03), pages 71-77, XP002087853 the whole document | 1-9, 11-13 |
| A | US 5 522 491 A (BAUDAT GASTON ET AL) 4 June 1996 (1996-06-04) abstract | 1,13 |

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

| Date of the actual completion of the international search | Date of mailing of the international search report |
|---|---|
| 5 April 2000 | 12/04/2000 |

| Name and mailing address of the ISA | Authorized officer |
|---|---|
| European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016 | Sonius, M |

1

# INTERNATIONAL SEARCH REPORT

Information on patent family members

Internal. J Application No
PCT/IB 99/02012

| Patent document cited in search report | Publication date | Patent family member(s) | Publication date |
|---|---|---|---|
| US 5522491 A | 04-06-1996 | CH 684222 A | 29-07-1994 |
| | | CA 2088349 A | 11-09-1993 |
| | | DE 59308678 D | 23-07-1998 |
| | | EP 0560023 A | 15-09-1993 |
| | | ES 2118142 T | 16-09-1998 |
| | | HK 1011834 A | 16-07-1999 |
| | | JP 6044377 A | 18-02-1994 |
| | | NO 930867 A | 13-09-1993 |
| | | US 5503262 A | 02-04-1996 |

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－73406

(43)公開日　平成11年(1999)３月16日

| (51)Int.Cl.⁶ | 識別記号 | ＦＩ | |
|---|---|---|---|
| Ｇ０６Ｆ　15/18 | ５６０ | Ｇ０６Ｆ　15/18 | ５６０Ａ |

審査請求　未請求　請求項の数11　ＯＬ　（全 13 頁）

(21)出願番号　特願平10－169787

(22)出願日　平成10年(1998)６月17日

(31)優先権主張番号　０８／８８３１９３

(32)優先日　1997年６月26日

(33)優先権主張国　米国（ＵＳ）

(71)出願人　596077259
ルーセント　テクノロジーズ　インコーポレイテッド
Ｌｕｃｅｎｔ　Ｔｅｃｈｎｏｌｏｇｉｅｓ　Ｉｎｃ．
アメリカ合衆国　07974　ニュージャージー、マレーヒル、マウンテン　アベニュー　600－700

(72)発明者　クリストファー　ジョン　バージェス
アメリカ合衆国，07728　ニュージャージー，フリーホールド，アンドラ　テラス　11

(74)代理人　弁理士　三俣　弘文

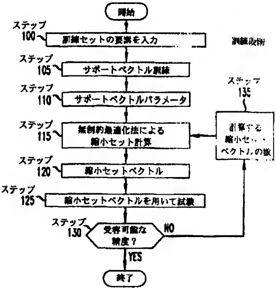
最終頁に続く

(54)【発明の名称】　サポートベクトル機械を使用する方法

(57)【要約】

【課題】　与えられたベクトルのセットを試験段階で用いるために高次元空間に写像するアルゴリズムを用いた機械の効率を改善する。

【解決手段】　サポートベクトル機械（ＳＶＭ）は、その判定面がサポートベクトルのセットによって、および、対応する重みのセットによって、パラメトライズされる万能学習機械である。本発明によるＳＶＭは、縮小セットベクトルを用いる。縮小セットベクトルの数は、セット内のベクトルの数より少ない。これらの縮小セットベクトルはセット内のベクトルとは異なり、同次２次核で用いられる固有値計算とは異なる最適化法に従って決定される。実施例では、パターン認識で用いるために、ＳＶＭは縮小セットベクトルを利用し、これにより、ユーザが選択するファクタだけこのＳＶＭの効率を改善する。これらの縮小セットベクトルは、無制約最適化法に従って決定される。



【特許請求の範囲】

【請求項1】　入力データ信号を受け取るステップと、前記入力データ信号に作用可能なサポートベクトル機械を用いて出力信号を生成するステップとからなる、サポートベクトル機械を使用する方法において、

前記サポートベクトル機械は縮小セットベクトルを利用し、

前記縮小セットベクトルは、同次2次核に用いられる固有値計算以外の最適化法を用いて訓練段階中にあらかじめ決定されたものであることを特徴とする、サポートベクトル機械を使用する方法。

【請求項2】　前記訓練段階は、

訓練セットの要素を受け取るステップと、

$N_s$個のサポートベクトルからなるサポートベクトルセットを生成するステップと、

$m \leqq N_s$として、縮小セットベクトルの数mを選択するステップと、

無制約最適化法を用いてm個の縮小セットベクトルを生成するステップとからなることを特徴とする請求項1に記載の方法。

【請求項3】　前記最適化法は無制約最適化法であることを特徴とする請求項1に記載の方法。

【請求項4】　前記入力データ信号は相異なるパターンを表し、前記出力信号は、該相異なるパターンの分類を表すことを特徴とする請求項1に記載の方法。

【請求項5】　前記訓練段階は、

前記サポートベクトル機械を訓練してサポートベクトルの数$N_s$を決定するステップと、

$m \leqq N_s$として、無制約最適化法を用いて、m個の縮小セットベクトルを決定するステップとからなることを特徴とする請求項1に記載の方法。

【請求項6】　入力データ信号を提供するデータ入力要素と、

前記入力データ信号に作用して少なくとも1つの出力信号を生成するサポートベクトル機械とからなる装置において、

前記サポートベクトル機械は、同次2次核に用いられる固有値計算以外の最適化法を用いてあらかじめ決定された縮小セットベクトルを用いて前記入力データ信号に作用することを特徴とする、サポートベクトル機械を用いた装置。

【請求項7】　前記データ入力要素は、該データ入力要素に入力された複数の画像を表す入力データ信号を提供することを特徴とする請求項6に記載の装置。

【請求項8】　前記少なくとも1つの出力信号は、各画像の分類を表すことを特徴とする請求項7に記載の装置。

【請求項9】　前記縮小セットベクトルの数はサポートベクトルの数より少ないことを特徴とする請求項6に記載の装置。

【請求項10】　前記最適化法は無制約最適化法であることを特徴とする請求項6に記載の装置。

【請求項11】　前記縮小セットベクトルは、前記無制約最適化法を用いて前記サポートベクトル機械を訓練している間にあらかじめ決定されることを特徴とする請求項10に記載の装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、万能学習機械に関し、特に、サポートベクトル機械に関する。

【0002】

【従来の技術】サポートベクトル機械（SVM(Support Vector Machine)）は、その判定面がサポートベクトルのセットによって、および、対応する重みのセットによって、パラメトライズされる万能学習機械である。SVMはまた、核関数によっても特徴づけられる。核の選択は、その結果として得られるSVMが多項式クラシファイアであるか、2層ニューラルネットワークであるか、動径（放射状）基底関数（RBF）機械であるか、あるいはその他の学習機械であるかを決定する。SVMの判定規則は、対応する各関数およびサポートベクトルの関数である。

【0003】

【発明が解決しようとする課題】一般に、SVMは、訓練段階および試験段階という2つの段階で動作する。訓練段階中、判定規則で用いるためのサポートベクトルのセットが生成される。試験段階中、特定の判定規則を用いて判定が行われる。残念ながら、この試験段階において、SVM判定規則の計算量は、サポートベクトルセット内のサポートベクトルの数$N_s$に比例する。

【0004】

【課題を解決するための手段】本発明によれば、与えられたベクトルのセットを試験段階で用いるために高次元空間に写像するアルゴリズムを用いた機械の効率を改善する方法および装置が実現される。具体的には、本発明の原理によれば、縮小セットベクトルを用いる。縮小セットベクトルの数は、セット内のベクトルの数より少ない。これらの縮小セットベクトルはセット内のベクトルとは異なり、同次2次核で用いられる固有値計算とは異なる最適化法に従って決定される。

【0005】本発明の実施例では、パターン認識で用いるために、SVMは縮小セットベクトルを利用し、これにより、ユーザが選択するファクタだけこのSVMの効率を改善する。これらの縮小セットベクトルは、無制約最適化法に従って決定される。

【0006】本発明の特徴によれば、縮小セットベクトルの選択により、性能対計算量のトレードオフを直接制御することが可能となる。

【0007】さらに、本発明の考え方はパターン認識に固有ではなく、サポートベクトルアルゴリズムが用いら

れるような任意の問題（例えば、回帰推定）に適用可能である。

【0008】

【発明の実施の形態】本発明の実施例について説明する前に、サポートベクトル機械について簡単な背景的知識を説明した後、本発明の考え方自体の説明を行う。本発明の考え方以外に、読者は、当業者に知られている核ベースの方法を一般的に表現するために用いられる数学的記法を知っていると仮定する。また、本発明の考え方は、パターン認識の場合の例に関して説明される。しかし、本発明の考え方は、サポートベクトルアルゴリズムが用いられるような任意の問題（例えば、回帰推定）に適用可能である。

【0009】以下の説明で、注意すべき点であるが、10個の数字のグレイレベル画像を含む2つの光学的文字認識（OCR）データセットからの試験データを用いる。一方のデータセットは、7，291個の訓練パターンおよび2，007個の試験パターンからなり、ここでは「郵便セット」という（例えば、L. Bottou, C. Cortes, H. Drucker, L. D. Jackel, Y. LeCun, U. A. Mueller, E. Saeckinger, P.Simard, and V. Vapnik, "Comparison of Classifier Methods: A Case Studyin Handwritten Digit Recognition", Proceedings of the 12th＊ IAPR International Conference on Pattern Recognition, Vol.2, IEEE Computer Society Press（米国カリフォルニア州ロスアラモス）、pp.77-83, 1994、および、Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D.Jackel, "Backpropagation Applied to Handwritten ZIP Code Recognition",Neural Computation, 1, 1989, pp.541-551、参照）。他方のデータセットは、NIST Special Database 3およびNIST Test Data 1からの60，000個の訓練パターンおよび10，000個の試験パターンからなり、ここでは「NISTセット」という（例えば、R. A. Wilkinson, J. Geist, S. Janet, P. J. Grother,C. J. C. Burges, R. Creecy, R. Hammond, J. J. Hull, N. J. Larsen, T. P.Vogl and C. L. Wilson, "The First Census Optical Character Recognition System Conference, 米国商務省, NIST, August 1992、参照）。「郵便セット」の画像は16×16ピクセルであり、「NISTセット」の画像は28×28ピクセルである。

【0010】［従来技術：サポートベクトル機械］判定規則が次の形をとるような2クラスクラシファイアを考える。

【数1】

$$y = \Theta\left(\sum_{i=1}^{N_S} \alpha_i K(\mathbf{x}, \mathbf{s}_i) + b\right) \tag{1}$$

ただし、ベクトルxおよびベクトル$s_i$は$R^d$の元であり、$\alpha_i$およびbは実数であり、Θは階段関数である。$R^d$はd次元ユークリッド空間であり、Rは実数である。$\alpha_i$、ベクトル$s_i$、$N_s$およびbはパラメータであり、ベクトルxは分類されるべきベクトルである。さまざまなクラシファイアに対する判定規則がこの形で書ける。例えば、K（x・$s_i$）$^p$は多項式クラシファイアを実現し、

【数2】

$$K = \exp(-\|\mathbf{x} - \mathbf{s}_i\|^2 / \sigma^2)$$

は動径基底関数機械を実現し、K＝tanh（γ（x・$s_i$）＋δ）は2層ニューラルネットワークを実現する（例えば、V. Vapnik, "Estimation of Dependencies Based on Empirical Data", Springer Verlag, 1982、V. Vapnik, "The Nature of Statistical Learning Theory", Springer Verlag, 1995、Boser, B. E., Guyon, I. M., and Vapnik, V., "A training algorithm foroptimal margin classifiers", Fifth Annual Workshop on Computational Learning Theory, Pittsburgh ACM 144-152, 1992、およびB. Schoelkopf, C. J. C. Burges, and V. Vapnik, "Extracting Support Data for a Given Task", Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press（米国カリフォルニア州Menlo Park, 1995、参照）。

【0011】サポートベクトルアルゴリズムは、その判定規則が式（1）の形をとるような任意の学習機械を訓練する原理的な方法である。要求される唯一の条件は、核Kが一般的な正値性制約を満たすことである（例えば、前掲の"The Nature of Statistical Learning Theory"、および"A training algorithm for optimal margin classifiers"、参照）。他の方法とは異なり、SVM訓練プロセスは、パラメータセット全体｛$\alpha_i$、ベクトル$s_i$、$N_s$およびb｝を決定する。その結果得られるベクトル$s_i$（i＝1，...，$N_s$）は、訓練セットのサブセットであり、サポートベクトルと呼ばれる。

【0012】サポートベクトル機械はいくつかの優れた性質を有する。訓練手続きは、制約2次最適化問題を解くこととなり、従って、求められる解は、目的関数の一意的な大域的最小値であることが保証される。SVMは、構造的リスク最小化を直接に実現するために使用可能である。この場合、学習機械の容量は、汎化誤りの限界を最小にするように制御することができる（例えば、前掲の"The Nature of Statistical Learning Theory"、および"Extracting Support Data for a GivenTask"、参照）。サポートベクトル判定面は、実際には、高次元空間内の線形分離超平面である。同様に、SVM

は、回帰を構成するためにも使用可能であり、これはある高次元空間において線形である（例えば、前掲の"The Nature of Statistical Learning Theory"、参照）。

【0013】サポートベクトル学習機械は、光学的文字認識（OCR）（例えば、前掲の"The Nature of Statistical Learning Theory"、および"Extracting Support Data for a Given Task"、ならびにC. Cortes and V. Vapnik, "Support Vector Networks", Machine Learning, Vol.20, pp.1-25, 1995、参照）、および対象認識のようなパターン認識問題に適用されて成功している。

【0014】図1は、従来技術のSVMの動作の流れ図である。この動作は、訓練段階および試験段階という2つの段階からなる。訓練段階では、ステップ52で、SVMは、クラスがあらかじめ割り当てられた訓練セットの要素を受け取る。ステップ54で、訓練セットからの入力データベクトルを多次元空間内へ変換する。ステップ56で、最適な多次元超平面に対するパラメータ（すなわち、サポートベクトルおよび対応する重み）が決定される。

【0015】図2に、訓練データ要素が2つのクラスに分離される例を示す。一方のクラスは円で表され、他方のクラスは四角で表されている。これは典型的な2クラスパターン認識問題のものである。例えば、「車」のパターンを「車でない」パターンから分離するように訓練されたSVMである。最適超平面は、2つのクラスのベクトルの間に極大マージンを有する線形判定関数である。すなわち、最適超平面は、訓練データを極大マージンで分離する一意的な判定面である。図2に示すように、最適超平面は、2つのクラスの間の分離が最大である領域によって定義される。図2で観察されるように、最適超平面を構成するには、訓練されたデータ要素のうち、この極大マージンを決定するサブセットを考慮すればよい。訓練要素のうち、最適超平面のパラメータを決定するこのサブセットは、サポートベクトルとして知られている。図2では、サポートベクトルは網掛けで示されている。

【0016】最適超平面は、高次元空間における写像されたサポートベクトルの線形結合で表される。SVMアルゴリズムは、ベクトルのセットに関する誤差が、すべてのサポートベクトルに重みを割り当てることによって最小化されることを保証する。これらの重みは、サポートベクトルよって判定面を計算する際に用いられる。また、このアルゴリズムによれば、特定の問題に属する訓練データに関する誤り率を最小にするために、これらの重みを適応させることが可能となる。これらの重みは、SVMの訓練段階中に計算される。

【0017】このようにして、最適超平面を構成することは、訓練セットの要素および写像された空間内の内積を決定する関数によって決定される制約2次最適化計画問題になる。この最適化問題に対する解は、従来の中間最適化法を用いて求められる。

【0018】一般に、最適超平面は、誤りなしで訓練データを分離することを必要とする。しかし、場合によっては、訓練データは誤りなしで分離することができないことがある。このような場合、SVMは、最小数の誤りで訓練データを分離しようと試み、残りの要素を極大マージンで分離する。このような超平面は一般に、ソフトマージン超平面として知られている。

【0019】試験段階では、ステップ62で、SVMは、分類すべき試験セットの要素を受け取る。次に、SVMは、サポートベクトルを核のパラメータとして用いて、試験セットの入力データベクトルを多次元空間に写像することによって変換する（ステップ64）。写像関数は、SVMにあらかじめロードされている核の選択によって決定される。この写像は、1つのベクトルをとり、それを高次元特徴空間へ変換して、線形判定関数がこの高次元特徴空間に生成されるようにする。図1の流れ図は陰(implicit)の写像を示しているが、この写像は陽(explicit)に実行されることも可能である。ステップ66で、SVMは、各入力データベクトルの所属状態を示すように、判定面から分類信号を生成する。最終結果は、図2に示されるように、円の（＋1）および四角の（－1）という出力分類信号の生成である。

【0020】残念ながら、式（1）の計算量は、サポートベクトルの数$N_s$に比例する。サポートベクトルの数の期待値は（l－1）E［P］でおさえられる。ただし、Pは、与えられたSVMを1個の訓練サンプルで訓練した場合の、1つの試験ベクトルに対する誤りの確率であり、E［P］は、l個のサンプルのすべての選び方にわたるPの期待値である（例えば、前掲の"The Nature of Statistical Learning Theory"、参照）。従って、$N_s$はおよそlに比例することが予想される。実際のパターン認識問題では、この結果、同様の汎化性能を有する他のシステムよりも試験段階において大幅に遅い機械が得られる（例えば、前掲の"Comparisonof Classifier Methods: A Case Study in Handwritten Digit Recognition"、および、Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Mueller, E. Saeckinger, P. Simard, and V. Vapnik, "Comparison of Learning Algorithms for Handwritten Digit Recognition", International Conference on Artificial Neural Networks, Ed. F. Fogelman, P. Gallinari, pp.53-60, 1995、参照）。

【0021】［縮小セットベクトル］これに対して、本発明の原理によれば、ずっと少数の縮小セットベクトルによりSVM判定規則を近似する方法および装置が実現される。縮小セットベクトルは以下の性質を有する。

【0022】・縮小セットベクトルは、サポートベクトルが完全なSVM判定規則に現れるのと同様にして、近似的なSVM判定規則に現れる。

・縮小セットベクトルは、サポートベクトルではない。縮小セットベクトルは、サポートベクトルとは異なり、必ずしも分離マージン上にはなく、訓練サンプルでもない。

・縮小セットベクトルは、与えられた、訓練済みのSVMに対して計算される。

・縮小セットベクトルの数（従って、結果として得られるSVMの試験段階における速度）は事前に選択される。

・縮小セット法は、サポートベクトル法が用いられる場合であればどのような場合にも適用可能である（例えば、回帰推定）。

【0023】［縮小セット］訓練データは、Lの要素、ベクトルxであるとする。ただし、L（Lは低次元(low dimensional)の意味）は、$d_L$次元ユークリッド空間$R^{d_L}$として定義される。SVMは、陰写像

【数3】

$$\Phi : \mathbf{x} \mapsto \bar{\mathbf{x}}, \quad \bar{\mathbf{x}} \in H$$

を実行する。ただし、H（Hは高次元(high dimensional)の意味）＝$R_{d_H}$、$d_H \leqq \infty$である。以下では、Hのベクトルにはバーを付けて示す。写像Φは、核Kの選択によって決定される。実際、Mercerの正値性制約（例えば、前掲の"The Nature of Statistical Learning Theory"、および"A training algorithm for optimal margin classifiers"、参照）を満たす任意のKに対して、＊

【数4】

$$K(\mathbf{x}_i, \mathbf{x}_j) = \bar{\mathbf{x}}_i \cdot \bar{\mathbf{x}}_j$$

であるようなペア｛Φ, H｝が存在する。従って、Hにおいて、SVM判定規則は単に（上記のように）、線形分離超平面となる。写像Φは通常、陽に計算されず、Hの次元$d_H$は通常大きい（例えば、同次写像$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^p$に対して、

【数5】

$$d_H = \binom{p + d_L - 1}{p}$$

である（$p+d_L-1$個のものからp個のものを選ぶ場合の数。従って、4次多項式で、$d_L=256$の場合、$d_H$は約1.8億となる）。

【0024】基本的なSVMパターン認識アルゴリズムは、2クラス問題を解く（例えば、前掲のEstimation of Dependencies Based on Empirical Data"、"The Nature of Statistical Learning Theory"、および"A training algorithm for optimalmargin classifiers"、参照）。訓練データ$\mathbf{x} \in L$および対応するクラスラベル$y_i \in \{-1, 1\}$が与えられた場合、SVMアルゴリズムは、ベクトル$\mathbf{x}_i$（$i=1, \ldots, l$）を2つのクラスに分ける判定面バーΨ∈Hを次のように構成する。

【数6】

$$\bar{\Psi} \cdot \bar{\mathbf{x}}_i + b \geq k_0 - \xi_i, \quad y_i = +1 \qquad (2)$$

$$\bar{\Psi} \cdot \bar{\mathbf{x}}_i + b \leq k_1 + \xi_i, \quad y_i = -1 \qquad (3)$$

ただし、$\xi_i$は正のスラック変数であり、分離不能の場合（例えば、前掲の"Support Vector Networks"、参照）を扱うために導入したものである。分離可能の場合、SVMアルゴリズムは、Hにおける正と負の例の間のマージンが最大になるような分離超平面を構成する。その後、

【数7】

$$\bar{\Psi} \cdot \Phi(\mathbf{x}) + b$$

が$(k_0+k_1)/2$より大きいか小さいかに応じて、試験ベクトル$\mathbf{x} \in L$にクラスラベル｛＋1，－1｝を割り当てる。サポートベクトル$\mathbf{s} \in L$は、式（2）または（3）のいずれかが等式になるような訓練サンプルとし※

※て定義される。（サポートベクトルは、他の訓練データと区別するためにベクトルsと表す。）すると、バーΨは次のように与えられる。

【数8】

$$\bar{\Psi} = \sum_{a=1}^{N_S} \alpha_a y_a \Phi(\mathbf{s}_a) \qquad (4)$$

ただし、$\alpha_a \geqq 0$は、訓練中に決定される重みであり、$y_a \in \{-1, 1\}$は、ベクトル$\mathbf{s}_a$のクラスラベルであり、$N_S$はサポートベクトルの数である。こうして、試験点ベクトルxを分類するためには、次式を計算する。

【数9】

$$\bar{\Psi} \cdot \bar{\mathbf{x}} = \sum_{a=1}^{N_S} \alpha_a y_a \bar{\mathbf{s}}_a \cdot \bar{\mathbf{x}} = \sum_{a=1}^{N_S} \alpha_a y_a K(\mathbf{s}_a, \mathbf{x}) \qquad (5)$$

【0025】しかし、本発明の考え方により、ここで、

【数10】

$$\Psi' = \sum_{a=1}^{N_Z} \gamma_a \Phi(\mathbf{z}_a) \qquad (6)$$

が距離尺度

【数11】

$$\rho = \left\| \bar{\Psi} - \bar{\Psi}' \right\| \qquad (7)$$

を最小にする（固定した$N_Z$に対して）ようなセット、

ベクトル$z_a$∈L（a＝1，...，$N_z$）および対応する重み$\gamma_a$∈Rを考える。

【0026】ここで、{$\gamma_a$，$z_a$}（a＝1，...，＊

$$\bar{\Psi}' \cdot \bar{\mathbf{x}} = \sum_{a=1}^{N_z} \gamma_a \bar{\mathbf{z}}_a \cdot \bar{\mathbf{x}} = \sum_{a=1}^{N_z} \gamma_a K(\mathbf{z}_a, \mathbf{x}) \qquad (8)$$

【0027】すると、目標は、結果として得られる汎化性能の損失が許容可能な範囲にとどまるような、最小の$N_z$<<$N_s$、および対応する縮小セットを選択することである。明らかに、$N_z$＝$N_s$とすることにより、ρを0にすることができる。しかし、$N_z$<$N_s$で、しかもρ＝0であるような自明でない場合が存在する（後述）。そのような場合、縮小セットにより、汎化性能の損失なしで、判定規則の計算量が低減される。各$N_z$に対して、対応する縮小セットを計算する場合、ρは、$N_z$の単調減少関数と見ることが可能であり、汎化性能もまた$N_z$の関数となる。本明細書では、汎化性能の$N_z$依存性に関する経験的結果のみについて説明する。

【0028】写像Φについて、以下のことに注意すべきである。Φの像は一般に線形空間にはならない。また、Φは一般に全射にはならず、一対一でない可能性がある（例えば、Kが偶数次の同次多項式の場合）。さらに、Φは、L内の線形従属ベクトルをH内の線形独立ベクトルに写像することがあり得る（例えば、Kが非同次多項式の場合）。Kが同次多項式の場合であっても、一般に、ベクトル$z_a$をスケールすることによって係数$\gamma_a$を1にスケールすることはできない（例えば、Kが偶数次の同次式である場合、$\gamma_a$は{＋1，－1}にスケールすることは可能であるが、必ずしも1にスケールすることはできない）。

【0029】［厳密解］このセクションでは、ρの最小値を解析的に計算する問題を考える。まず、簡単ではあるが自明ではない場合について説明する。

【0030】［同次2次多項式］同次2次多項式の場合、規格化を1に選ぶ。

K（$x_i$，$x_j$）＝（$x_i$・$x_j$）$^2$　（9）

【0031】説明を簡単にするため、1次近似$N_z$＝1を計算する。対称テンソル

【数13】

$$S_{\mu\nu} \equiv \sum_{i=1}^{N_s} \alpha_i y_i s_{i\mu} s_{i\nu} \qquad (10)$$

を導入する。※

$$\rho^2 = S_{\mu\nu} S^{\mu\nu} - \sum_{a=1}^{N_z} \gamma_a^2 \|\mathbf{z}_a\|^4 \qquad (14)$$

ただし、固有ベクトルは、固有値の絶対値の大きさの順に並べるものとする。なお、trace（$S^2$）は、Sの固有値の平方の和であるので、$N_z$＝$d_L$（データの次元）と選択することにより、近似は厳密（すなわちρ＝0）になる。サポートベクトルの数$N_s$は$d_L$より大きいことが多いため、このことは、汎化性能の損失なしに、縮小セットのサイズはサポートベクトルの数より小さくなりうることを示している。

＊$N_z$）を縮小セットという。試験点ベクトルxを分類するには、式（5）の展開を次の近似で置き換える。

【数12】

※【数14】

$$\rho = \left\| \bar{\Psi} - \gamma \bar{\mathbf{z}} \right\|$$

は、次式を満たす{γ，ベクトルz}に対して最小になることが分かる。

【数15】

$$S_{\mu\nu} z_\nu = \gamma z^2 z_\mu \qquad (11)$$

（繰り返す添字については和をとる）。{γ，ベクトルz}をこのように選ぶと、$\rho^2$は次のようになる。

【数16】

$$\rho^2 = S_{\mu\nu} S^{\mu\nu} - \gamma^2 z^4 \qquad (12)$$

【0032】従って、{γ，ベクトルz}を、ベクトルzが、Sの固有値λ＝γ$z^2$が最大絶対値を有するような固有ベクトルとなるように選択するときに、ρの最大降下が達成される。なお、γは、γ＝sign{λ}となるように選択することが可能であり、ベクトルzは$z^2$＝｜λ｜となるようにスケールすることが可能である。

【0033】オーダー$N_z$に拡張すると、同様にして、式

【数17】

$$\rho = \left\| \bar{\Psi} - \sum_{a=1}^{N_z} \gamma_a \bar{\mathbf{z}}_a \right\| \qquad (13)$$

を最小にするセット{$\gamma_i$，ベクトル$z_i$}におけるベクトル$z_i$は、それぞれ固有値が

【数18】

$$\gamma_i \left\| \mathbf{z}_i \right\|^2$$

であるSの固有ベクトルであることが示される。これにより次式が成り立ち、ρの降下は、ベクトル$z_a$をSのはじめの$N_z$個の固有ベクトルに選択した場合に最大となる。

【数19】

【0034】一般の場合、縮小セットを計算するためには、ρは、すべての｛$y_a$，ベクトル$z_a$｝（$a=1, \ldots, N_Z$）にわたって同時に最小にならなければならない。次のような反復法を考えると便利である。すなわち、第iステップでは、｛$y_j$，ベクトル$z_j$｝（$j<i$）を固定して、｛$y_i$，ベクトル$z_i$｝を計算する。2次多項式の場合、この反復法によって生成される最小値の列が、問題全体に対する最小値も生成する。この結果は、2次多項式に特有であり、ベクトル$z_i$が直交する（あるいはそのように選択することができる）という事実の結果である。

【0035】以下の表1に、試験セットに関して誤りの数$E_Z$を達成するために必要な縮小セットサイズ$N_Z$を示す。ここで、郵便セットに関して訓練された2次多項式SVMの場合、$E_Z$は、サポートベクトルの完全セットを用いて求められる誤りの数$E_S$とは、高々1個の誤りしか異ならない。明らかに、2次の場合、縮小セットは、精度をほとんど失うことなく、計算量を大幅に減らすことができる。また、多くの数字では、サポートベクトルの数は$d_L=256$より大きいが、これは、精度を全く失わずに高速化が可能であることを示す。

【表1】

| | サポートベクトル | | 縮小セット | |
|---|---|---|---|---|
| 数字 | $N_S$ | $E_S$ | $N_Z$ | $E_Z$ |
| 0 | 292 | 15 | 10 | 16 |
| 1 | 95 | 9 | 6 | 9 |
| 2 | 415 | 28 | 22 | 29 |
| 3 | 403 | 26 | 14 | 27 |
| 4 | 375 | 35 | 14 | 34 |
| 5 | 421 | 26 | 18 | 27 |
| 6 | 261 | 13 | 12 | 14 |
| 7 | 228 | 18 | 10 | 19 |
| 8 | 446 | 33 | 24 | 33 |
| 9 | 330 | 20 | 20 | 21 |

＊【0036】［一般の核］縮小セット法を任意のサポートベクトル機械に適用するには、上記の解析を一般の核に拡張しなければならない。例えば、同次多項式K（$x_1$，$x_2$）＝N（$x_1 \cdot x_2$）$^n$の場合、反復法の最初のペア｛$y_1$，ベクトル$z_1$｝を求めるために

【数20】

$$\frac{\partial \rho}{\partial z_{1a_1}} = 0$$

とおくと、式（11）に類似の次式が得られる．

【数21】

＊

$$S_{\mu_1\mu_2\ldots\mu_n} z_{1\mu_2} z_{1\mu_3} \cdots z_{1\mu_n} = \gamma_1 \|\mathbf{z}_1\|^{2n-2} z_{1\mu_1} \quad (15)$$

ただし、※

※【数22】

$$S_{\mu_1\mu_2\ldots\mu_n} \equiv \sum_{m=1}^{N_S} \alpha_m y_m s_{m\mu_1} s_{m\mu_2} \cdots s_{m\mu_n} \quad (16)$$

である。

【0037】この場合、γに関してρを変化させても新しい条件は得られない。1次の解｛$y_1$，ベクトル$z_1$｝★

★に対して式（15）を解いたとすると、ρ′は次のようになる。

【数23】

$$\rho^2 = S_{\mu_1\mu_2\ldots\mu_n} S^{\mu_1\mu_2\ldots\mu_n} - \gamma_1^2 \|\mathbf{z}_1\|^{2n} \quad (17)$$

【0038】そこで、次のような定義をすることができる。☆

☆【数24】

$$\tilde{S}_{\mu_1\mu_2\ldots\mu_n} \equiv S_{\mu_1\mu_2\ldots\mu_n} - \gamma_1 z_{1\mu_1} z_{1\mu_2} \cdots z_{1\mu_n} \quad (18)$$

これにより、2次の解ベクトル$z_2$に対する反復方程式が、式（15）でS、ベクトル$z_1$および$y_1$をそれぞれ～S、ベクトル$z_2$および$y_2$で置き換えた形をとる。（なお、2より高い次数の多項式では、ベクトル$z_a$は一般に直交しない。）しかし、これらは反復解のみであり、さらに、すべての｛$y_a$、ベクトル$z_a$｝が同時に変化することを許容した場合の連立方程式を解く必要がある。さらに、これらの方程式は複数の解を有し、そのほとんどはρに関する極小値に対応する。さらに、別のKを選択することにより、他の固定点方程式が得られる。式（15）の解は反復（すなわち、任意のベクトルzから始めて、式（15）を用いて新たなベクトルzを計算し、これを繰り返す）によって求められるが、次のセクションで説明する方法はさらに柔軟で強力である。

【0039】［無制約最適化法］核Kの1次導関数が定義されていると仮定すると、未知数｛$y_i$、ベクトル$z_i$｝に関する目的関数$F \equiv \rho^2/2$の勾配を計算することができる。例えば、K（$s_a$、$s_b$）がスカラー$s_a \cdot s_b$の関数であると仮定すると、次式のようになる。

【数25】

$$\frac{\partial F}{\partial \gamma_k} = -\sum_{m=1}^{N_S} \alpha_m y_m K(\mathbf{s}_m \cdot \mathbf{z}_k) + \sum_{j=1}^{N_z} \gamma_j K(\mathbf{z}_j \cdot \mathbf{z}_k) \qquad (19)$$

$$\frac{\partial F}{\partial z_{k\mu}} = -\sum_{m=1}^{N_S} \gamma_k \alpha_m y_m K'(\mathbf{s}_m \cdot \mathbf{z}_k) s_{m\mu} + \sum_{j=1}^{N_z} \gamma_j \gamma_k K'(\mathbf{z}_j \cdot \mathbf{z}_k) z_{j\mu} \qquad (20)$$

【0040】従って、本発明の原理によれば、（おそらくは局所的な）最小は、無制約最適化法を用いて求めることができる。

【0041】［アルゴリズム］まず、所望の近似次数$N_z$を選択する。$X_i \equiv$ ｛$y_i$、$z_i$｝とする。2段階法を用いる。第1段階（後述）で、すべてのベクトル$z_j$（j＜i）を固定したまま、$X_i$を反復的に計算する。

【0042】第2段階（後述）で、すべての$X_i$が変動することを許容する。

【0043】注意すべき点であるが、式（20）における勾配は、$\gamma_k$が0である場合、0である。この事実は、重大な数値的不安定性につながる可能性がある。この問題を回避するために、第1段階は、単純な「レベル交差」定理に基づく。そのアルゴリズムは以下のとおりである。まず、$\gamma_i$を＋1または－1に初期化し、ベクトル$z_i$をランダム値で初期化する。次に、$\gamma_i$を固定したままで、ベクトル$z_i$を変化させる。次に、ベクトル$z_i$、$X_j$（j＜i）を固定した場合の、$\gamma_i$の最適値を解析的に計算する。次に、ベクトル$z_i$および$\gamma_i$の両方に関して同時にFを最小化する。最後に、すべてのj≦iに対して最適な$\gamma_j$を解析的に計算する。これは、$\Gamma = Z^{-1}\Delta$によって与えられる。ただし、δ、ΓおよびZは次のように与えられる（式（19）参照）。

【数26】

$$\Gamma_j \equiv \gamma_j \qquad (21)$$

$$\Delta_j \equiv \sum_{a=1}^{N_S} \alpha_a y_a K(\mathbf{s}_a \cdot \mathbf{z}_j) \qquad (22)$$

$$Z_{jk} \equiv K(\mathbf{z}_j, \mathbf{z}_k) \qquad (23)$$

【0044】Zは正定値かつ対称であるため、周知のコレスキー分解を用いて効率的に逆行列を求めることができる。

【0045】こうして、アルゴリズムの第1段階は以下のように進行する。

［1］$\gamma_1$＝＋1または－1をランダムに選び、ベクトル$z_1$をランダム値に設定する。

［2］ベクトル$z_1$を変化させてFを最小化する。

［3］ベクトル$z_1$を固定したまま、Fをさらに最大に低下させる$\gamma_1$を計算する。

［4］ベクトル$z_1$、$\gamma_1$をともに変化させてさらにFを低下させる。

［5］最良の解を保持してステップ［1］～［4］をT回反復する。

［6］ベクトル$z_1$、$\gamma_1$を固定し、$\gamma_2$＝＋1または－1をランダムに選び、ベクトル$z_2$をランダム値に設定する。

［7］ベクトル$z_2$を変化させてFを最小化する。

［8］ベクトル$z_2$（およびベクトル$z_1$、$\gamma_1$）を固定し、Fをさらに最大に低下させる最適な$\gamma_2$を計算する。

［9］｛ベクトル$z_2$、$\gamma_2$｝をともに変化させてさらにFを低下させる。

［10］最良の解を保持してステップ［6］～［9］をT回反復する。

［11］最後に、ベクトル$z_1$、ベクトル$z_2$を固定し、（上記の式（21）～（23）に示されるように）さらにFを低下させる最適な$\gamma_1$、$\gamma_2$を計算する。

【0046】次に、この手続きを｛ベクトル$z_3$、$\gamma_3$｝、｛ベクトル$z_4$、$\gamma_4$｝など、

【数27】

$$\{\mathbf{z}_{N_Z} : \gamma_{N_Z}\}$$

まで反復する。

【0047】$y_i$が0に近づかないようにすることによって数値的不安定性は回避される。上記のアルゴリズムにより、これは自動的に保証される。第1ステップで、$y_i$を固定したままベクトル$z_i$を変化させた結果、目的関数Fが減少した場合、次に$y_i$を変化させるときに、$y_i$は0を通ることはできない。その理由は、0を通るとすると（その場合｛ベクトル$z_i$、$y_i$｝のFへの寄与は0となるので）Fが増大してしまうからである。

【0048】なお、与えられた｛ベクトル$z_i$、$y_i$｝のペアの各計算は、第1段階で、ベクトル$X_i$に対する相異なる初期値で数回（T回）反復される。Tは、求められたFにおける相異なる最小値の個数Mから経験的に決定される。上記のデータセットでは、Mは通常2または3であり、TはT＝10と選ばれた。

【0049】第2段階では、第1段階で求められたすべてのベクトル$X_i$が単一のベクトルへと連接され、すべてのパラメータの変動を許容して、再び無制約最小化プロセスが適用される。注意すべき点であるが、第2段階の結果、目的関数Fがさらに約2分の1に減少することが多い。

【0050】本発明の原理に従って、以下の1次無制約最適化法を両方の段階で用いた。探索方向は、共役勾配法を用いて求められる。探索方向に沿って、ブラケット点$x_1$，$x_2$および$x_3$を、$F(x_1)>F(x_2)<F(x_3)$となるように求める。次に、このブラケットを平衡化する（平衡化法については、例えば、W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, "Numerical Recipes in C", Second Edition, Cambridge University Press, 1992、参照）。これらの3点を通る2次当てはめ曲線の最小値を、次の反復で選択される開始点として用いる。共役勾配プロセスは、所定の反復数の後に再開され、全体のプロセスは、Fの減少率があるしきい値を下回ったときに終了する。注意すべき点であるが、この一般的アプローチは、上記の2次多項式核の場合に適用した場合、解析的アプローチと同じ結果を与えた。

【0051】［実験］上記のアプローチを、郵便セットに対して最良の性能を有するSVMに適用した。このSVMは次数3の非同次多項式機械（これについては、例えば、前掲の"The Nature of Statistical Learning Theory"、参照）であった。近似の次数$N_z$は、各2クラスクラシファイアに対して試験段階で10倍の高速化がなされるように選択した。結果を表2（下記）に示す。縮小セット法は、精度の損失はほとんどなしで、この高速化を達成した。10個のクラシファイアをまとめて1つの10クラスクラシファイア（これについては、例えば、前掲の"The Natureof Statistical Learning Theory"、および"Support Vector Networks"、参照）として用いると、完全サポートセット（サポートセット全体）を用いた場合には4．2％のエラーであるのに対して、縮小セットを用いた場合には4．3％のエラーであった。なお、組み合わせた場合、縮小セットでは6倍の高速化しか得られない。その理由は、相異なる2クラスクラシファイアがいくつかの共通のサポートベクトルを有し、キャッシングの可能性があるためである。これらの方法をさらに大きい問題に拡張することができるかどうかという問題を解決するため、NISTセットの場合に、数字0を他のすべての数字から分離する2クラスクラシファイアに対して研究を繰り返した（60，000回の訓練、10，000個の試験パターン）。このクラシファイアも、完全サポートセットを用いて、最良の精度を与えるもの（次数4の多項式）を選んだ。1，273個のサポートベクトルの完全セットでは19個の試験エラーを生じたが、サイズ127の縮小セットでは20個の試験エラーであった。

【0052】

【表2】

| | サポートベクトル | | 縮小セット | |
|---|---|---|---|---|
| 数字 | $N_S$ | $E_S$ | $N_Z$ | $E_Z$ |
| 0 | 272 | 13 | 27 | 13 |
| 1 | 109 | 9 | 11 | 10 |
| 2 | 380 | 26 | 38 | 26 |
| 3 | 418 | 20 | 42 | 20 |
| 4 | 392 | 34 | 39 | 32 |
| 5 | 397 | 21 | 40 | 22 |
| 6 | 257 | 11 | 26 | 11 |
| 7 | 214 | 14 | 21 | 13 |
| 8 | 463 | 26 | 46 | 28 |
| 9 | 387 | 13 | 39 | 13 |
| 計 | 3289 | 187 | 329 | 188 |

【0053】（なお、試験は、完全な10桁のNISTに対しても行われ、10％の精度損失で50倍の高速化がなされた。C. J. C. Burges, B. Schoelkopf, "Improvingthe Accuracy and Speed of Support Vector Machines", in press, NIPS '96、参照。）

【0054】

【実施例】図3に、SVMの訓練段階で用いられる、本発明の原理を実現する例示的な流れ図を示す。ステップ100で、入力訓練データがSVM（図示せず）に入力される。ステップ105で、SVMがこの入力データに対して訓練され、ステップ110で、SVMはサポートベクトルのセットを生成する。ステップ135で、縮小セットベクトルの数が選択される。ステップ115で、無制約最適化法（前述）を用い、ステップ120で縮小セットベクトルを生成する。ステップ125で、この縮小セットベクトルを用いて、サンプルデータのセット（図示せず）を試験する。ステップ130で、この試験の結果を評価する。試験結果が（例えば速度および精度に関して）受容可能な場合、この縮小セットベクトルが

以後利用される。試験結果が受容可能でない場合、縮小セットベクトルを決定するプロセスを再び実行する。（後者の場合、注意すべき点であるが、（例えば速度あるいは精度に関する）試験結果は、縮小セットベクトルの数をさらに少なくすることを示唆する可能性もある。）

【0055】縮小セットベクトルが決定されると、SVMで利用可能となる。この縮小セットベクトルを試験段階で使用する方法を図4に示す。ステップ215で、試験セットからの入力データベクトルがSVMに送られる。ステップ220で、SVMは、縮小セットベクトルを核のパラメータとして用いて、試験セットの入力データベクトルを多次元空間に写像することにより変換する。ステップ225で、SVMは、判定面から、各入力データベクトルの帰属状態を示す分類信号を生成する。

【0056】上記のように、m個の縮小セットベクトルが縮小セット内にある。これらの縮小セットベクトルは、図3に示した上記の訓練段階で決定される。速度および精度のデータが、m個より少ない縮小セットベクトルを使用することも可能であることを示唆する場合、別のアプローチを用いて、新たなさらに小さい縮小セットベクトルのセットを再計算する必要を回避することが可能である。特に、x＜mとして、x個の縮小セットベクトルは、m個の縮小セットベクトルのセットから選択される。この場合、いくつ（x）の縮小セットベクトルを使用するかの決定は、例えば、訓練段階で生成された速度および精度のデータを用いて経験的に行われる。しかし、これらの縮小セットベクトルの値を再計算する必要はない。

【0057】パターン認識の場合の、本発明の考え方の実施例を図5に示す。パターン認識システム100は、プロセッサ105および認識器110からなり、認識器110は、データ入力要素115、およびSVM120からなる。本発明の考え方以外には、図5の要素は周知であるため、詳細には説明しない。例えば、データ入力要素115は、分類するための入力データをSVM120へ送る。データ入力要素115の一例はスキャナである。この場合、入力データは画像のピクセル表現（図示せず）である。SVM120は、本発明の原理に従って縮小セットベクトルを用いて入力データに作用する。動作（試験）中、SVM120は、入力データの分類を表す数値結果を、後続の処理のためにプロセッサ105に送る。プロセッサ105は、例えば、メモリを伴うマイクロプロセッサのような蓄積プログラム制御プロセッサである。プロセッサ105は、さらに、例えば自動預払機（ATM）などにおける認識器110の出力信号を処理する。

【0058】図5のシステムは2つのモード、すなわち、訓練モードおよび動作（試験）モードで動作する。訓練モードの例は、図3に示される上記の方法である。試験モードの例は、図4に示される上記の方法である。

【0059】以上、本発明について説明したが、当業者には認識されるように、本発明の技術的範囲内でさまざまな変形例を考えることができる。例えば、本発明の考え方は、サポートベクトル機械以外の、核に基づく方法にも適用可能であり、例えば、回帰推定、密度評価などにも使用可能であるが、これらに限定されるものではない。

【0060】

【発明の効果】以上述べたごとく、本発明によれば、与えられたベクトルのセットを試験段階で用いるために高次元空間に写像するアルゴリズムを用いた機械の効率を改善する方法および装置が実現される。本発明の特徴によれば、縮小セットベクトルの選択により、性能対計算量のトレードオフを直接制御することが可能となる。さらに、本発明の考え方はパターン認識に固有ではなく、サポートベクトルアルゴリズムが用いられるような任意の問題（例えば、回帰推定）に適用可能である。

【図面の簡単な説明】

【図1】従来技術のSVMの動作の流れ図である。

【図2】代表サポートベクトルにより訓練データを2つのクラスに分離する一般的な図である。

【図3】本発明の原理に従ってSVMシステムを訓練する例示的な方法の図である。

【図4】本発明の原理に従ってSVMシステムを動作させる例示的な方法の図である。

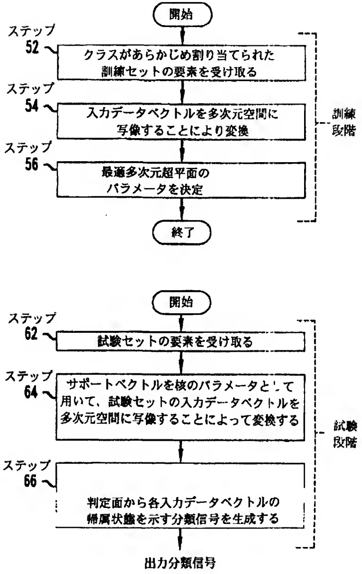
【図5】本発明の原理を実現する認識システムの一部のブロック図である。

【符号の説明】

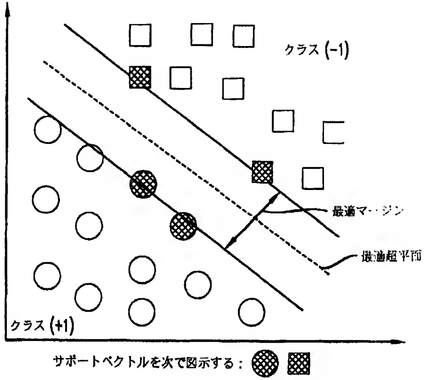
100 パターン認識システム
105 プロセッサ
110 認識器
115 データ入力要素
120 SVM

【図1】

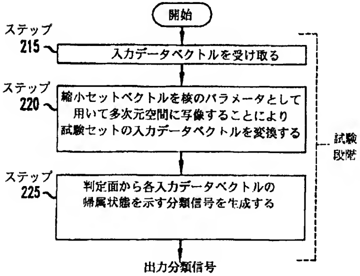
（従来技術）


開始
ステップ 52
クラスがあらかじめ割り当てられた訓練セットの要素を受け取る
ステップ 54
入力データベクトルを多次元空間に写像することにより変換
ステップ 56
最適多次元超平面のパラメータを決定
訓練段階
終了
開始
ステップ 62
試験セットの要素を受け取る
ステップ 64
サポートベクトルを核のパラメータとして用いて、試験セットの入力データベクトルを多次元空間に写像することによって変換する
ステップ 66
判定面から各入力データベクトルの帰属状態を示す分類信号を生成する
試験段階
出力分類信号

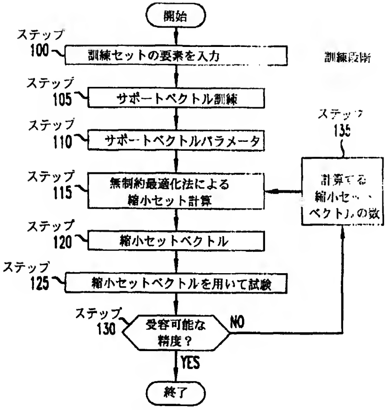
【図2】


クラス (-1)
最適マージン
最適超平面
クラス (+1)
サポートベクトルを次で図示する：

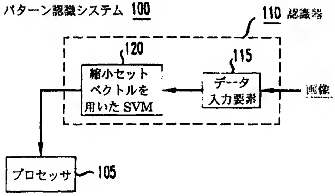
【図4】


開始
ステップ 215
入力データベクトルを受け取る
ステップ 220
縮小セットベクトルを核のパラメータとして用いて多次元空間に写像することにより試験セットの入力データベクトルを変換する
ステップ 225
判定面から各入力データベクトルの帰属状態を示す分類信号を生成する
試験段階
出力分類信号

【図3】




【図5】




フロントページの続き

(71)出願人 596077259
600 Mountain Avenue, Murray Hill, New Jersey 07974－0636U. S. A.